APPLICATION OF SMIRNOV WORDS TO WAITING TIME DISTRIBUTIONS OF RUNS

UTA FREIBERG, CLEMENS HEUBERGER, AND HELMUT PRODINGER

Abstract. Consider infinite random words over a finite alphabet where the letters occur as an i.i.d. sequence according to some arbitrary distribution on the alphabet. The expectation and the variance of the waiting time for the first completed h-run of any letter (i.e., first occurrence of h subsequential equal letters) is computed.

The expected waiting time for the completion of h-runs of j arbitrary distinct letters is also given.

1. Introduction

In [7], the following paradox is presented: In measuring the regularity of a die one may use waiting times for sequences of the same side of certain lengths. For example, if ones throws a regular six-sided die, it takes 7 throws on average to get a number subsequently twice and 43 throws to get a number three times in succession. Heuristically, one would expect that a *smaller* number of throws is needed to get such sequences with a *biased* die. This leads to the definition to call one die *more regular* than another one if more throws are needed to get sequences of one side of a certain length. Now the paradox is that there exist dice—say A and B—where the mean waiting time for two digits in a row is longer for die A while the mean waiting time for three digits in a row is longer for die B (an example has been given by Móri, see [7, p. 62]). The consequence of this paradox is that one cannot use the mean waiting times for such runs as a (sufficient) criterion for the definition of regularity of a die (or whatever random sequence of digits from a finite alphabet).

This paradox gave motivation to calculate first and second moments of such waiting times for so called *h-runs*. In particular, the formula

2010 *Mathematics Subject Classification.* 05A05; 05A15, 60C05, 60G40.

Key words and phrases. Waiting time distribution, run, Smirnov word, generating function.

Parts of the article were written while Uta Freiberg was a visitor at Stellenbosch University.

Clemens Heuberger is supported by the Austrian Science Fund (FWF): P 24644-N26. Parts of the article were written while Clemens Heuberger was a visitor at Stellenbosch University.

Helmut Prodinger is supported by an incentive grant of the NRF of South Africa.

for the first moment of the waiting time for the first completed h-run of *any* digit—which was already given in [7]—is proved without using the strong law of large numbers or any other limit theorem (see Theorem 1). Moreover, the variance of the waiting time for the first completed h-run is presented in the same theorem. We then compute the waiting time for the completion of h-runs of j different letters in Theorem 2. In particular, for $j = r$ (the number of possible letters), we get results about the waiting time for a full collection of runs.

Our fundamental technique is the calculation of generating functions of such waiting times; our main trick is the combination of two very useful observations: Firstly, we make use of the very simple but crucial identity (1) (see [1]) which already has been a powerful tool in the treatment of the coupon collector problem and/or the birthday paradox. Secondly, we use the generating function of Smirnov words (see [2]) to count words with a limited number of repetitions of single letters using an appropriate substitution.

We conclude the paper in Section 5 with an algorithmic approach for specific situations.

2. Preliminaries

We consider infinite words $X_1X_2\dots$ over the alphabet $\mathcal{A} = \{1, \dots, r\}$ where the random variables X_i are i.i.d. with $\mathbb{P}\{X_i = k\} = p_k > 0$ for some $p_1, \dots, p_r$.

We say that a letter $\ell \in \mathcal{A}$ has an *h-run in $X_1 \dots X_n$* if there are h consecutive letters ℓ in the word $X_1 \dots X_n$, or in other words, if the word $\ell^h = \ell\ell \dots \ell$ (with h repetitions) is a factor of the word $X_1 \dots X_n$.

We consider the random variable B_j giving the first position n such that there exist j of the r letters having an h-run in $X_1 \dots X_n$. This is a random variable on the infinite product space consisting of all infinite words endowed with the product measure.

On the other hand, we consider the random variable Y_n counting the number of letters which had an h-run in $X_1 \dots X_n$. This is a random variable on the finite product space consisting of all words of length n, again with its product measure.

By construction, we have

$$\mathbb{P}\{Y_n \geq j\} = \mathbb{P}\{B_j \leq n\}, \tag{1}$$

cf. [1, Eqn. (6)]. As a consequence, we obtain (cf. [1, Eqn. (7)])

$$\mathbb{E}(B_j) = \sum_{n\geq 0} \mathbb{P}\{B_j > n\} = \sum_{n\geq 0} \mathbb{P}\{Y_n < j\} = \sum_{q=0}^{j-1} \sum_{n\geq 0} \mathbb{P}\{Y_n = q\}. \tag{2}$$

With the generating function

$$G_j(z) = \sum_{n\geq 0} \mathbb{P}\{Y_n < j\} z^n, \tag{3}$$

this amounts to

$$\mathbb{E}(B_j) = G_j(1).$$

To compute the variance, we first note that

$$\begin{aligned}\mathbb{E}(B_j^2) &= \sum_{n\geq 0} n^2 \mathbb{P}\{B_j = n\} = \sum_{n\geq 0} n^2 (\mathbb{P}\{B_j > n-1\} - \mathbb{P}\{B_j > n\}) \\ &= \sum_{n\geq 0} (n+1)^2 \mathbb{P}\{B_j > n\} - \sum_{n\geq 0} n^2 \mathbb{P}\{B_j > n\} \\ &= \sum_{n\geq 0} (2n+1)\mathbb{P}\{B_j > n\} = \sum_{n\geq 0} (2n+1)\mathbb{P}\{Y_n < j\} \\ &= 2G_j'(1) + G_j(1)\end{aligned}$$

where we used (1) and the definition of $G_j(z)$ given in (3). We conclude that

$$\mathbb{V}(B_j) = \mathbb{E}(B_j^2) - \mathbb{E}(B_j)^2 = 2G_j'(1) + G_j(1) - G_j(1)^2. \tag{4}$$

A *Smirnov word* is defined to be any word which has no consecutive equal letters. The ordinary generating function of Smirnov words over the alphabet $\mathcal{A}$ is

$$S(v_1, \ldots, v_r) = \frac{1}{1 - \sum_{i=1}^{r} \frac{v_i}{1+v_i}} \tag{5}$$

where v_i counts the number of occurrences of the letter i, cf. Flajolet and Sedgewick [2, Example III.24].

3. Moments of the first h-run

In this section, we study the first occurrence of any h-run. In the framework of Section 2, this corresponds to the case $j = 1$ and the random variable B_1.

We prove the following result on the expectation of B_1:

Theorem 1. *If $p_i < 1$ for $1 \leq i \leq r$, the expectation and the variance of the first occurrence of an h-run are*

$$\mathbb{E}(B_1) = \frac{1}{\sum_{i=1}^{r} \frac{1}{p_i^{-1} + \cdots + p_i^{-h}}} \tag{6}$$

and

$$\mathbb{V}(B_1) = \frac{\sum_{i=1}^{r} \left(\frac{p_i + p_i^h}{1 - p_i^h} - 2h \frac{p_i^h (1-p_i)}{(1-p_i^h)^2} \right)}{\left(\sum_{i=1}^{r} \frac{1}{p_i^{-1} + \cdots + p_i^{-h}} \right)^2}. \tag{7}$$

The result (6) on the expectation also appears (without proof) in [7, p. 62]. Each summand of the numerator of (7) is indeed non-negative, because this is equivalent to

$$\frac{p_i + p_i^h}{2} \cdot \frac{1 + p_i + \cdots + p_i^{h-1}}{h} \geq p_i^h,$$

which is true by the inequality between the arithmetic and the geometric mean, applied to both factors.

Proof of Theorem 1. In the case $j = 1$, (2) reads

$$\mathbb{E}(B_1) = \sum_{n\geq 0} \mathbb{P}\{Y_n = 0\}. \tag{8}$$

Thus we have to determine the probability that a word of length n does not have any h-run. Such words arise from a Smirnov word by replacing single letters by runs of length in $\{1, \dots, h-1\}$ of the same letter.

In terms of generating function, this corresponds to replacing each v_i by

$$p_i z + \cdots + (p_i z)^{h-1} = \frac{p_i z - (p_i z)^h}{1 - p_i z}.$$

Here, z marks the length of the word. We obtain

$$\begin{aligned} G_1(z) = \sum_{n\geq 0} \mathbb{P}\{Y_n = 0\} z^n &= S\left(\frac{p_1 z - (p_1 z)^h}{1 - p_1 z}, \dots, \frac{p_r z - (p_r z)^h}{1 - p_r z}\right) \\ &= \frac{1}{1 - \sum_{i=1}^r \frac{\frac{p_i z - (p_i z)^h}{1 - p_i z}}{1 + \frac{p_i z - (p_i z)^h}{1 - p_i z}}} = \frac{1}{1 - \sum_{i=1}^r \frac{p_i z - (p_i z)^h}{1 - (p_i z)^h}}. \end{aligned}$$

By (8), we are only interested in $z = 1$:

$$\mathbb{E}(B_1) = \sum_{n\geq 0} \mathbb{P}\{Y_n = 0\} = G_1(1) = \frac{1}{1 - \sum_{i=1}^r \frac{p_i - p_i^h}{1 - p_i^h}}.$$

Replacing the summand 1 in the denominator by $p_1 + \cdots + p_r$ yields

$$\begin{aligned} \mathbb{E}(B_1) &= \frac{1}{\sum_{i=1}^r \left(p_i - \frac{p_i - p_i^h}{1 - p_i^h}\right)} = \frac{1}{\sum_{i=1}^r \frac{p_i - p_i^{h+1} - p_i + p_i^h}{1 - p_i^h}} \\ &= \frac{1}{\sum_{i=1}^r \frac{p_i^h (1 - p_i)}{1 - p_i^h}} = \frac{1}{\sum_{i=1}^r \frac{1}{p_i^{-1} + \cdots + p_i^{-h}}}. \end{aligned}$$

For the variance, we compute $G_1'(1)$ as

$$G_1'(1) = \mathbb{E}(B_1)^2 \sum_{i=1}^r \frac{(p_i - h p_i^h)(1 - p_i^h) + (p_i - p_i^h) h p_i^h}{(1 - p_i^h)^2}$$

$$= \mathbb{E}(B_1)^2 \sum_{i=1}^{r} \frac{p_i - hp_i^h - p_i^{h+1} + hp_i^{2h} + hp_i^{h+1} - hp_i^{2h}}{(1-p_i^h)^2}$$
$$= \mathbb{E}(B_1)^2 \sum_{i=1}^{r} \frac{p_i(1-p_i^h) - hp_i^h(1-p_i)}{(1-p_i^h)^2}$$
$$= \mathbb{E}(B_1)^2 \biggl(\sum_{i=1}^{r} \frac{p_i}{1-p_i^h} - h \sum_{i=1}^{r} \frac{p_i^h(1-p_i)}{(1-p_i^h)^2} \biggr).$$

By (4), we obtain

$$\begin{aligned}
\mathbb{V}(B_1) &= 2G_1'(1) + G_1(1) - G_1(1)^2 \\
&= \mathbb{E}(B_1)^2 \biggl(-1 + 2 \sum_{i=1}^{r} \frac{p_i}{1-p_i^h} - 2h \sum_{i=1}^{r} \frac{p_i^h(1-p_i)}{(1-p_i^h)^2} \\
&\qquad\qquad + \sum_{i=1}^{r} \frac{p_i^h(1-p_i)}{1-p_i^h} \biggr) \\
&= \mathbb{E}(B_1)^2 \biggl(\sum_{i=1}^{r} \frac{-p_i + p_i^{h+1} + 2p_i + p_i^h - p_i^{h+1}}{1-p_i^h} \\
&\qquad\qquad - 2h \sum_{i=1}^{r} \frac{p_i^h(1-p_i)}{(1-p_i^h)^2} \biggr) \\
&= \mathbb{E}(B_1)^2 \biggl(\sum_{i=1}^{r} \frac{p_i + p_i^h}{1-p_i^h} - 2h \sum_{i=1}^{r} \frac{p_i^h(1-p_i)}{(1-p_i^h)^2} \biggr).
\end{aligned}$$

Together with (6), we obtain (7). □

4. Expectation of the first occurrence of h-runs of j letters

In this section, we consider the first position where j of the letters $1, \ldots, r$ had an h-run. In the terminology of Section 2, this corresponds to the random variable B_j.

We prove the following theorem on the expectation of B_j.

Theorem 2. *For $i \in \mathcal{A}$, let*

$$\alpha_i := \frac{p_i - p_i^h}{1-p_i}, \qquad \gamma_i := \frac{p_i}{1-p_i} \tag{9}$$

and let A_i and Γ_i be the substitution operators mapping the variable v_i to α_i and γ_i, respectively.

Then the expectation of the first occurrence of h-runs of exactly j letters is

$$\mathbb{E}(B_j) = \biggl(\sum_{q=0}^{j-1} [y^q] \prod_{i=1}^{r} (y\Gamma_i + (1-y)A_i) \biggr) S(v_1, \ldots, v_r), \tag{10}$$

where $S(v_1, \dots, v_r)$ is defined in (5).

For $j = r$, i.e., the first occurrence of h-runs of all letters, (10) can be simplified:

Corollary 3. *The expectation of the first occurrence of all h-runs is*

$$\mathbb{E}(B_r) = \left(\prod_{i=1}^{r} \Gamma_i - \prod_{i=1}^{r} (\Gamma_i - A_i)\right) S(v_1, \dots, v_r), \tag{11}$$

where Γ_i, A_i and $S(v_1, \dots, v_r)$ are defined in (9) *and* (5)*, respectively.*

In the case of equidistributed letters, i.e., $p_i = 1/r$ for all i, we get the following simple expression.

Corollary 4. *If $p_1 = \cdots = p_r = 1/r$, then the expectation of the first occurrence of all h-runs is*

$$\mathbb{E}(B_r) = \frac{r(r^h - 1)}{r - 1} H_r,$$

where H_r denotes the rth harmonic number.

Proof of Theorem 2. As in Section 2, Y_n is the number of letters that have at least one run of length $\geq h$ within $X_1 \dots X_n$.

Arbitrary words arise from Smirnov words by replacing single letters by runs of length at least 1 of the same letter. In terms of generating functions, this corresponds to substituting v_i by

$$\begin{aligned} &p_i z + \cdots + (p_i z)^{h-1} + u_i((p_i z)^h + (p_i z)^{h+1} + \cdots) \\ &\quad = \frac{p_i z - (p_i z)^h + u_i (p_i z)^h}{1 - p_i z} = \frac{p_i z + (u_i - 1)(p_i z)^h}{1 - p_i z} =: \beta_i(u_i, z). \end{aligned}$$

As previously, z counts the length of the word. The variable u_i counts the number of occurrences of (non-extensible) m-runs of the letter i with $m \geq h$.

We now consider the probability generating function

$$F(u_1, \dots, u_r; z) = S(\beta_1(u_1, z), \dots, \beta_r(u_r, z)).$$

of all words.

For $M \subseteq \mathcal{A}$, let $E_{n,M}$ be the event that exactly the letters in M have an h-run in $X_1 \dots X_n$. By definition, we have

$$\{Y_n = q\} = \biguplus_{\substack{M \subseteq \mathcal{A} \\ |M| = q}} E_{n,M} \tag{12}$$

for $q \in \{0, \dots, r\}$.

We now compute $\mathbb{P}(E_{n,M})$ for some $M = \{i_1, \dots, i_q\}$ of cardinality q. We denote the letters not contained in M by $\mathcal{A} \setminus M = \{s_1, \dots, s_{n-q}\}$.

By construction of the generating function, we have

(13)
$$\mathbb{P}(E_{n,M}) = [z^n][u_{s_1}^0]\cdots[u_{s_{n-q}}^0] \sum_{m_{i_1},\dots,m_{i_q}\geq 0} [u_{i_1}^{m_{i_1}}]\cdots[u_{i_q}^{m_{i_q}}]F(u_1,\dots,u_r;z).$$

For any power series $H(u)$, we have

$$\sum_{m\geq 1}[u^m]H(u) = H(1) - H(0).$$

We therefore define the operators Δ_i and Z_i by $\Delta_i H(u_i) = H(1)-H(0)$ and $Z_i H(u_i) = H(0)$. With these notations, (13) reads

$$\mathbb{P}(E_{n,M}) = [z^n]\Big(\prod_{i\in M}\Delta_i \prod_{i\notin M} Z_i\Big)F(u_1,\dots,u_r;z). \tag{14}$$

Inserting this and (12) in (2) yields

$$\mathbb{E}(B_j) = \sum_{n\geq 0}[z^n] \sum_{\substack{M\subseteq\mathcal{A}\\ |M|<j}} \Big(\prod_{i\in M}\Delta_i \prod_{i\notin M} Z_i\Big)F(u_1,\dots,u_r;z). \tag{15}$$

Summing over all $n \geq 0$ amounts to setting $z = 1$ as long as all summands are non-singular at $z = 1$. As $|M| < j$, at least one of the u_i is zero, w.l.o.g. $u_1 = 0$. This implies that $[z^n]F(u_1,\dots,u_r;z) \leq [z^n]F(0,1,\dots,1;z) < \rho^n$ for a suitable $0 < \rho < 1$ as the word 1^h is forbidden as a factor. Thus $F(u_1,\dots,u_r;z)$ is regular at $z = 1$.

We note that $\beta_i(1,1) = \gamma_i$ and $\beta_i(0,1) = \alpha_i$ where γ_i and α_i are defined in (9). Therefore, for $z = 1$, the operator Δ_i can be written as $\Gamma_i - A_i$. Similarly, Z_i corresponds to A_i.

We have

$$\sum_{\substack{M\subseteq\mathcal{A}\\ |M|<j}} \prod_{i\in M}(\Gamma_i - A_i)\prod_{i\notin M} A_i = \sum_{q=0}^{j-1}[y^q]\prod_{i=1}^{r}(y\Gamma_i + (1-y)A_i).$$

Combining this with (15) yields (10). □

Proof of Corollary 3. The polynomial $\prod_{i=1}^r(y\Gamma_i+(1-y)A_i)$ has degree r in the variable y. Thus extracting all coefficients but the coefficient of y^r amounts to substituting $y = 1$ and subtracting the coefficient of y^r, i.e.,

$$\sum_{q=0}^{j-1}[y^q]\prod_{i=1}^{r}(y\Gamma_i + (1-y)A_i) = \prod_{i=1}^{r}\Gamma_i - \prod_{i=1}^{r}(\Gamma_i - A_i).$$

Inserting this into (10) yields (11). □

Proof of Corollary 4. Setting $p_i = 1/r$ yields

$$\gamma_i = \frac{\frac{1}{r}}{1-\frac{1}{r}} = \frac{1}{r-1}, \qquad \alpha_i = \frac{\frac{1}{r} - \left(\frac{1}{r}\right)^h}{1-\frac{1}{r}} = \frac{1-\frac{1}{r^{h-1}}}{r-1},$$
$$\frac{\gamma_i}{1+\gamma_i} = \frac{1}{r}, \qquad \frac{\alpha_i}{1+\alpha_i} = \frac{r^{h-1}-1}{r^h-1}.$$

Inserting this in (11) and collecting terms with k occurrences of A_i yields

$$\begin{aligned}\mathbb{E}(B_r) &= \sum_{k=1}^{r} \binom{r}{k} (-1)^{k+1} \frac{1}{1-\frac{r-k}{r} - k\frac{r^{h-1}-1}{r^h-1}} \\ &= \frac{r(r^h-1)}{r-1} \sum_{k=1}^{r} \binom{r}{k} (-1)^{k+1} \frac{1}{k} = \frac{r(r^h-1)}{r-1} H_r,\end{aligned}$$

where we used the identity

$$H_r = \sum_{k=1}^{r} \binom{r}{k} \frac{(-1)^{k+1}}{k},$$

cf. [5]. □

Remark 5. Let run lengths $h_1, \dots, h_r$ be given and consider occurrences of h_i-runs for the letter i. If B_j is the first position n such that there are exactly j letters which had "their" run in $X_1 \dots X_n$, the results of Theorems 1 and 2 as well as Corollary 3 remain valid when all p_i^h are replaced by $p_i^{h_i}$.

5. Algorithmic Aspects

For fixed h, the occurrence of an h-run of the variable X_i can easily be detected by a transducer automaton reading the occurrence probabilities p_i and outputting 1 whenever the letter i completes an h run, see Figure 1 for the case $r = 2$, $h = 3$ and $i = 2$.

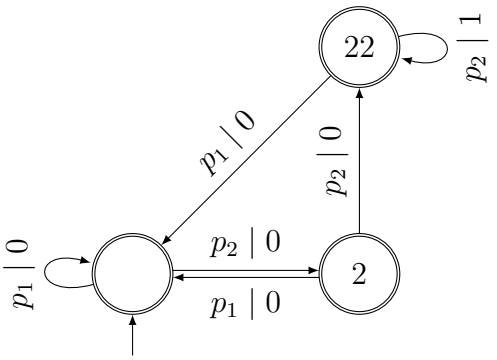

Figure 1. Transducer detecting 3-runs of the letter 1.

The same can be done for the first occurrence of any h-run, see Figure 2 for $r = 2$ and $h = 3$.

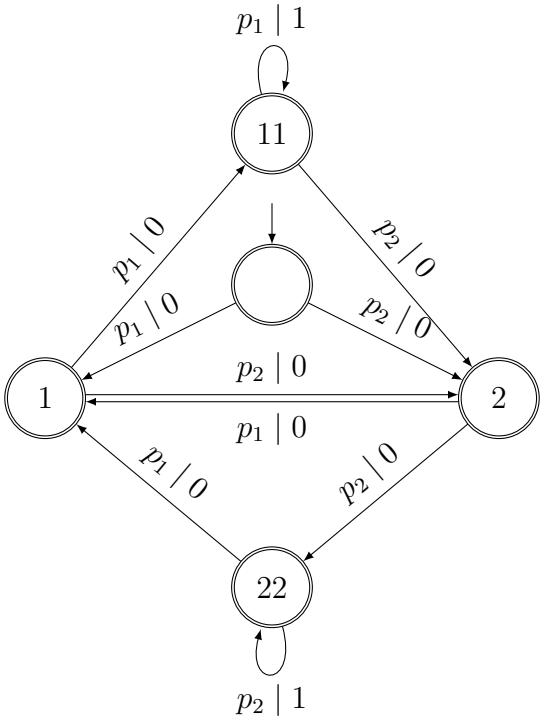

FIGURE 2. Transducer detecting the first 3-run of any letter.

The first occurrence of j runs of length h could also be modelled by a transducer.

Using the finite state machine package [4] of the SageMath Mathematics Software [6], such transducers can easily be constructed.

Accompanying this article, in [3], an extension of SageMath to compute the expectation and the variance of the first occurrence of a 1 in the output of a transducer is proposed for inclusion into SageMath.

Using this extension, the expectation and the variance of B_1 can be computed for fixed r and h as shown in Table 1.

The results coincide with those obtained in Theorem 1. For more examples, see the documentation of `moments_waiting_time`.

For $j > 1$, we did not compute $\mathbb{V}(B_j)$ in general. For fixed r and h, it can be computed by this algorithmic approach.

Obviously, the SageMath method can be used for computing first occurrences of everything which is recognisable by a transducer. On the other hand, explicit results for general r and h such as our Theorems 1 and 2 cannot be obtained by that method.

```
from sage.combinat import finite_state_machine as FSM

# Deactivate deprecated code
FSM.FSMOldCodeTransducerCartesianProduct = False
FSM.FSMOldProcessOutput = False

# Construct the polynomial ring and set up q
R.<p> = QQ[]
q = 1 - p

# Construct the Transducers detecting runs of single
# letters.  [p, p, p] is the block to detect, [p, q]
# the alphabet
p_runs = transducers.CountSubblockOccurrences(
    [p, p, p], [p, q])
q_runs = transducers.CountSubblockOccurrences(
    [q, q, q], [p, q])

# In order to detect runs of both letters, build the
# cartesian product ...
both_runs = p_runs.cartesian_product(q_runs)
# ... and add up the output by concatenating with
# the predefined "add" transducer on the alphabet
# [0, 1] We use the Python convention that any
# non-zero integer evaluates to True in boolean
# context.
first_run = transducers.add([0, 1])(both_runs)

# Declare it as a Markov chain
first_run.on_duplicate_transition = \
    FSM.duplicate_transition_add_input
print first_run.moments_waiting_time()
```

TABLE 1. Computation of the moments for B_1 with $r = 2$ and $h = 3$ in SageMath.

REFERENCES

[1] Philippe Flajolet, Danièle Gardy, and Loÿs Thimonier, *Birthday paradox, coupon collectors, caching algorithms and self-organizing search*, Discrete Appl. Math. **39** (1992), no. 3, 207–229.

[2] Philippe Flajolet and Robert Sedgewick, *Analytic combinatorics*, Cambridge University Press, Cambridge, 2009.

[3] Clemens Heuberger, *FiniteStateMachine: Moments of waiting time*, `http://trac.sagemath.org/ticket/18070`, 2015.

[4] Clemens Heuberger, Daniel Krenn, and Sara Kropf, *Automata and transducers in the computer algebra system Sage*, 2014, arXiv:1404.7458 [cs.FL].

[5] Peter J. Larcombe, Eric J. Fennessey, Wolfram A. Koepf, and David R. French, *On Gould's identity No. 1.45*, Util. Math. **64** (2003), 19–24.
[6] William A. Stein et al., *Sage Mathematics Software (Version 6.5)*, The Sage Development Team, 2015, `http://www.sagemath.org`.
[7] Gábor J. Székely, *Paradoxes in probability theory and mathematical statistics*, Mathematics and its Applications (East European Series), vol. 15, D. Reidel Publishing Co., Dordrecht, 1986, Translated from the Hungarian by Márta Alpár and Éva Unger.

Institut für Stochastik und Anwendungen, Universität Stuttgart, Pfaffenwaldring 57, D-70569 Stuttgart, Germany
E-mail address: `uta.freiberg@mathematik.uni-stuttgart.de`

Institut für Mathematik, Alpen-Adria-Universität Klagenfurt, Universitätsstrasse 65–67, 9020 Klagenfurt, Austria
E-mail address: `clemens.heuberger@aau.at`

Department of Mathematical Sciences, Stellenbosch University, 7602 Stellenbosch, South Africa
E-mail address: `hproding@sun.ac.za`